immersionED

AI Measurement Validation

XQ Synthesis Inter-Rater Reliability · Claude Opus 4.7 · v9 Production · April 2026

About this study. immersionED builds AI-powered adaptive simulations where students practice reasoning, communication, and decision-making in real-world scenarios. A core challenge is measuring complex student thinking reliably and at scale — without adding assessment burden to teachers. This study, developed in collaboration with XQ Institute, validates immersionED's AI scoring engine against expert human teachers on the XQ Synthesis competency framework. The goal is not to replace teacher judgment but to give educators a tool they can trust — one that scores the kinds of thinking that matter most with the same reliability as trained experts, so teachers can focus on instruction, feedback, and the work that only humans can do.

Claude Opus 4.7 scores XQ Synthesis transcripts with reliability that meets — and slightly exceeds — that of trained human teachers.

Cohen's quadratic weighted kappa against teacher consensus, on the Articles of Confederation State Delegate transcripts.

0.883	0.856	p=.003	11/15
AI VS. CONSENSUS	**HUMAN VS. LOO CONSENSUS**	**BLIND REVIEW**	**CELLS STABLE**
v9 · generic XQ only	3 expert teachers	AI preferred · 95% of cells	across 6 independent runs

OPTIMIZATION EFFECT · Same rubric, same model, same transcripts · Prompt optimization is the only variable

0.692	>>	0.883	+28%
unoptimized		v9 production	improvement

FOUR INDEPENDENT LINES OF EVIDENCE

01 Non-Inferior Reliability
k=0.883 vs human 0.856 · Non-inferiority at d=0.075, p=0.021

02 Preferred in Blind Review
4.23 vs 3.81 accuracy · p=0.003 · At-or-above human on 95% of cells

03 No Contextualization Needed
Generic XQ rubric alone = k=0.883 · AI self-contextualizes from simulation knowledge

04 No Training Data Required
Zero improvement from N=3-5 exemplars · Prompt optimization alone sufficient

Finding 1: Inter-Rater Reliability

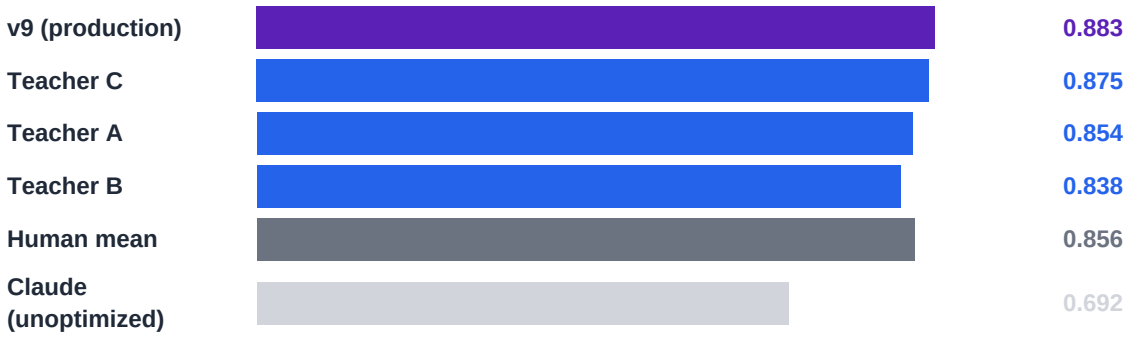
v9 production prompt vs 3 expert teachers · Cohen's quadratic weighted kappa against post-sync consensus

0.883
V9 KAPPA
Mean of 6 runs · sd=0.025

p=.021
NON-INFERIORITY
d=0.075 margin

p=.005
K > 0.8
Near-perfect threshold

KAPPA VS CONSENSUS — v9 OUTPERFORMS EVERY TEACHER



PER-CELL DELTA (lower = more aligned with consensus)

Rater	Kappa	Per-cell |D|	Stable Cells
v9 (production)	**0.883**	0.389	11 / 15
Teacher C	0.875	0.533	—
Teacher A	0.854	0.467	—
Teacher B	0.838	0.533	—
Human mean	0.856	0.511	—
Claude (unoptimized)	0.692	0.889	—

SIGNIFICANCE TESTS

Test	p-value
k > 0.8 threshold	**p=0.005**
Non-inferiority d=0.075	**p=0.021**
Non-inferiority d=0.10	**p=0.008**

Finding 2: Blind Peer Review

3 expert teachers rated blind-labeled scoring rationales · n=39 cells · 312 total ratings

4.23
AI ACCURACY
vs human 3.81 · p=0.003

3.15
AI EVIDENCE
vs human 2.00 · p<0.001

95%
WIN RATE
37 of 39 cells

ACCURACY RATINGS BY SOURCE (1-5 scale)

WIN / TIE / LOSS (n=39 cells)

11
AI SOLE WINNER
of 39

26
TIED AT TOP
of 39

2
HUMAN SOLE WINNER
of 39

OPTIMIZATION EFFECT

3.64 → 4.23 Claude original → v8 · Same rubric inputs, same model. Prompt optimization is the only variable.

SIGNIFICANCE

Comparison	D	p-value	Significant?
v8 vs human mean (accuracy)	+0.42	0.003	**Yes**
v8 vs human mean (evidence)	+1.15	<0.001	**Yes**
v8 vs Teacher A individually	+0.60	0.021	**Yes**
v8 vs Teacher C individually	+0.38	0.036	**Yes**
v8 vs Teacher B individually	+0.25	0.12	Directional

"All of them require assumptions and leaps of faith... without knowing the student or hearing their line of reasoning, the same transcript can be interpreted both ways depending on bias and perspective." — Expert teacher, blind review commentary

Finding 3: Rubric Isolation — The Path to v9

Does the AI need contextualized rubrics? Three configurations, 6 runs each.

Configuration	k mean	k sd	D mean	Stable Cells
A — v8 + both rubrics	0.867	0.030	6.50	7 / 15
B — generic XQ only = v9	**0.883**	**0.025**	**5.83**	**11 / 15**
C — contextualized only	0.855	0.038	6.50	9 / 15

All three statistically equivalent (pairwise bootstrap, all p > 0.15). Config B (v9) is best on every metric.

KAPPA BY CONFIGURATION

NOTABLE CELL-LEVEL EFFECTS

T10 UR — CEILING FIX
Consensus: L4 · Runs reaching L4:

Generic (v9): **6/6**
Both rubrics: 2/6
Context only: 1/6

T8 AC — KNOWN LIMITATION
Consensus: L0 · Runs reaching L0:

Context only: **6/6**
Both rubrics: 2/6
Generic (v9): 1/6

Why Generic-Only Works

The AI built the simulation. It already knows the scenario, characters, and debate structure. The contextualized rubric was humans spelling out what the AI already knows. Removing it eliminates interference and produces sharper judgments.

Scalability: A new simulation launches with just generic XQ progressions + v9 optimization layer. Zero per-course setup required.

Questions about this study?
Contact Chad Wilson, Founder & CEO · chad@immersioned.org · immersioned.org